

This listing of claims will replace all prior versions,
and listings, of claims in the application:

Claims 1-45 (deleted)

1 Claim 46 (currently amended): A method for filtering a
2 plurality of candidate search results to remove
3 near-duplicates, the method comprising:
4 a) ~~for each of a plurality of~~ one of the plurality of
5 candidate search results, determining whether the one
6 candidate search result is a near-duplicate of another
7 of the plurality of candidate search results ~~result~~ by
8 1) comparing a cluster identifier of the one
9 candidate search result with ~~that a cluster~~
10 identifier of the other candidate search result,
11 and
12 2) if the cluster identifiers of the one and the
13 other ~~two~~ candidate search results match, then
14 concluding that the one ~~two~~ candidate search is a
15 near-duplicate of the other candidate search
16 result ~~results are near duplicates~~; and
17 b) ~~if it is determined in response to a determination~~
18 that the one candidate search result is a
19 near-duplicate of the other candidate search result,
20 ~~then~~ rejecting the one candidate search result thereby
21 defining a filtered set of search results including
22 only those of the plurality of candidate search
23 results that have not been rejected.

1 Claim 47 (currently amended): A search filter for
2 processing a plurality of search results to remove
3 near-duplicates, the search filter comprising:
4 a) a near-duplicate determination facility for
5 determining, for ~~each of a plurality of~~ one of the
6 plurality of candidate search results, whether the one
7 candidate search result is a near-duplicate of another
8 of the plurality of candidate search results result,
9 and wherein the near-duplicate determination facility
10 includes a comparison facility for comparing a cluster
11 identifier of the one candidate search result with a
12 cluster identifier ~~that of another~~ the other candidate
13 search result, and wherein if the cluster identifiers
14 of the one ~~two~~ candidate search ~~results~~ result and the
15 other candidate search result match, then it is
16 concluded that the one ~~two~~ candidate search ~~results~~
17 result and the other candidate search result are
18 near-duplicates; and
19 b) a filter for rejecting the one candidate search
20 result if it is determined that the one candidate
21 search result is a near-duplicate of the other
22 candidate search result and passing the one candidate
23 search result if it is not determined that the one
24 candidate search result is a near-duplicate of the
25 other candidate search result, thereby defining a
26 filtered set of search results including only those of
27 the plurality of candidate results that have not been
28 rejected by the filter.

1 Claim 48 (previously presented): A machine-readable medium
2 having stored thereon a plurality of records, each of the
3 records comprising:

4 a) a first field for storing a document identifier;
5 and
6 b) a plurality of lists, each of the plurality of
7 lists containing elements of a document identified by
8 the document identifier stored in the first field,
9 wherein a hash function is used to hash each of
10 the elements in order to determine which one of the
11 plurality of lists that each of the elements will be
12 contained in.

1 Claim 49 (currently amended): A method for determining
2 whether two documents are near-duplicates, the method
3 comprising:
4 a) for each of the two documents, generating at least
5 two different fingerprints; and
6 b) determining whether or not the two documents are
7 near-duplicate documents by
8 1) determining whether or not any one of the at
9 least two fingerprints of a first of the two
10 documents matches any one of the at least two
11 fingerprints of a second of the two documents,
12 and
13 2) if it is determined that any one fingerprint
14 of the at least two fingerprints of the first of
15 the two documents does match any one fingerprint
16 of the at least two fingerprints of the second of
17 the two documents, then concluding that the two
18 documents are near-duplicates; and
19 c) using the determination of whether or not the
20 two documents are near-duplicates in at least one
21 of (A) an act of serving search results
22 corresponding to documents, (B) an act of

23 crawling documents, (C) an act of indexing
24 documents, and (D) an act of fixing a broken link
25 to at least one of the two documents.

1 Claim 50 (previously presented): A machine-readable medium
2 having stored thereon a plurality of records, each of the
3 records comprising:

- 4 a) a first field for storing a document identifier;
5 and
6 b) a plurality of lists, each of the plurality of
7 lists containing elements of a document identified by
8 the document identifier stored in the first field,
9 wherein at least some of the plurality of lists
10 include different numbers of elements.

1 Claim 51 (currently amended) The machine-readable medium
2 of claim ~~51~~ 50 wherein at least one of the plurality of
3 lists include no elements.

1 Claim 52 (currently amended): A machine-readable medium
2 having stored thereon a plurality of records, each of the
3 records comprising:

- 4 a) a first field for storing a document identifier;
5 and
6 b) a plurality of lists, each of the plurality of
7 lists containing elements of a document identified by
8 the document identifier stored in the first field,
9 wherein at least some contiguous elements in a
10 document are not necessarily contiguous elements of a list.

1 Claim 53 (previously presented): A machine-readable medium
2 having stored thereon a plurality of records, each of the
3 records comprising:

4 a) a first field for storing a document identifier;
5 and
6 b) a plurality of lists, each of the plurality of
7 lists containing elements of a document identified by
8 the document identifier stored in the first field,
9 wherein for each of the records, the number of
10 lists is the same.

1 Claim 54 (previously presented): The machine-readable
2 medium of claim 53 wherein a number of the plurality of
3 lists is independent of document size.

1 Claim 55 (previously presented): The machine-readable
2 medium of claim 48 wherein each of the elements of a
3 document is an element that has been extracted from the
4 document.

1 Claim 56 (previously presented): The machine-readable
2 medium of claim 48 wherein each of the elements of a
3 document is a predetermined one of (A) a predetermined
4 number of words, (B) a predetermined number of sentences,
5 (C) a predetermined number of characters, (D) a
6 predetermined number of paragraphs, and (E) a predetermined
7 number of sections.

1 Claim 57 (previously presented): The machine-readable
2 medium of claim 48 wherein each of the elements of a
3 document partially overlaps another of the elements of the
4 document.

1 Claim 58 (previously presented): The machine-readable
2 medium of claim 50 wherein each of the elements of a
3 document is an element that has been extracted from the
4 document.

1 Claim 59 (previously presented): The machine-readable
2 medium of claim 50 wherein each of the elements of a
3 document is a predetermined one of (A) a predetermined
4 number of words, (B) a predetermined number of sentences,
5 (C) a predetermined number of characters, (D) a
6 predetermined number of paragraphs, and (E) a predetermined
7 number of sections.

1 Claim 60 (previously presented): The machine-readable
2 medium of claim 50 wherein each of the elements of a
3 document partially overlaps another of the elements of the
4 document.

1 Claim 61 (previously presented): The machine-readable
2 medium of claim 52 wherein each of the elements of a
3 document is an element that has been extracted from the
4 document.

1 Claim 62 (previously presented): The machine-readable
2 medium of claim 52 wherein each of the elements of a
3 document is a predetermined one of (A) a predetermined
4 number of words, (B) a predetermined number of sentences,
5 (C) a predetermined number of characters, (D) a
6 predetermined number of paragraphs, and (E) a predetermined
7 number of sections.

1 Claim 63 (previously presented): The machine-readable
2 medium of claim 52 wherein each of the elements of a
3 document partially overlaps another of the elements of the
4 document.

1 Claim 64 (previously presented): The machine-readable
2 medium of claim 53 wherein each of the elements of a
3 document is an element that has been extracted from the
4 document.

1 Claim 65 (previously presented): The machine-readable
2 medium of claim 53 wherein each of the elements of a
3 document is a predetermined one of (A) a predetermined
4 number of words, (B) a predetermined number of sentences,
5 (C) a predetermined number of characters, (D) a
6 predetermined number of paragraphs, and (E) a predetermined
7 number of sections.

1 Claim 66 (previously presented): The machine-readable
2 medium of claim 53 wherein each of the elements of a
3 document partially overlaps another of the elements of the
4 document.

1 Claim 67 (previously presented): A machine-readable medium
2 having stored thereon a plurality of records, each of the
3 records comprising:
4 a) a first field for storing a document identifier;
5 and
6 b) a plurality of lists, each of the plurality of
7 lists containing elements of a document identified by
8 the document identifier stored in the first field,
9 wherein each of the elements are contained in one of

10 the plurality of lists in accordance with a result of
11 hashing the element using a hash function.